Using Depth and Appearance Features for Informed Robot Grasping of Highly Wrinkled Clothes

Arnau Ramisa, Guillem Alenyà, Francesc Moreno-Noguer and Carme Torras

***Abstract*— Detecting grasping points is a key problem in cloth manipulation. Most current approaches follow a multiple re-grasp strategy for this purpose, in which clothes are sequentially grasped from different points until one of them yields to a desired configuration. In this paper, by contrast, we circumvent the need for multiple re-graspings by building a robust detector that identifies the grasping points, generally in one single step, even when clothes are highly wrinkled.**

In order to handle the large variability a deformed cloth may have, we build a Bag of Features based detector that combines appearance and 3D geometry features. An image is scanned using a sliding window with a linear classifier, and the candidate windows are refined using a non-linear SVM and a "grasp goodness" criterion to select the best grasping point.

We demonstrate our approach detecting collars in deformed polo shirts, using a Kinect camera. Experimental results show a good performance of the proposed method not only in identifying the same trained textile object part under severe deformations and occlusions, but also the corresponding part in other clothes, exhibiting a degree of generalization.

Fig. 1. Robot executing a grasp on the detected collar in a scenario with several clothes.

I. Introduction

Manipulating textile objects is becoming a very active research topic due to its interest for service robotics and the availability of new, dexterous manipulation tools. However, in current research, the perception component of the task is usually disregarded or simplified as much as possible.

A complete system for retrieving one by one all elements of a laundry basket or pile, classifying and then folding them is proposed in [1]. In this approach, the topmost element of the pile is found using stereo vision, and its geometric center is used as a grasping point. The grasping operation is repeated as many times as necessary to ensure a correct grasp. Once the cloth object is held by the manipulator, four basic visual features are extracted and used in a 1-Nearest Neighbor classifier to select the corresponding learned object.

More related to our work, in [2] the authors use state-of-the-art computer vision techniques for the manipulation of socks. Local Binary Patterns and 2D shape features are used in a χ^2 Support Vector Machine (SVM) with an extended Gaussian kernel to determine the sock type and if the sock is inside-out. A model-based approach is used to determine sock configuration for posterior manipulation and pairing with a PR2 robot.

This work was supported by the Spanish Ministry of Science and Innovation under projects PAU+ DPI2011-27510, Consolider projects MIPRCV CSD2007-00018 and AT CSD2007-0022, the EU Project IntellAct FP7-ICT2009-6-269959 and by AGAUR through SGR-00155 and SGR-1434.

The authors are with the Institut de Robotica i Informatica Industrial (CSIC-UPC), Llorens i Artigas 4-6, 08028 Barcelona, Spain {aramisa,galenya,fmoreno,torras}iri.upc.edu

In [3] the authors designed a cloth grasping-point selection system to autonomously take elements from a pile of washed towels and fold and stack them using a Willow Garage PR2 robot. The method uses vision to detect corners that can be used for re-grasping the towel when it is already hanging from one of the robot arms. The initial pick-up is done by selecting the central point of the cloth, detected through background segmentation and stereo.

In [4] the authors describe a complete system, designed for the PR2 robot, for laundry handling. The system starts by picking up an unknown piece of clothing, identifying it, and then bringing it to a desired configuration. For the first task, namely, identification and estimation of the current state, the cloth item is initially grasped by an arbitrary edge, and a series of low-hanging-point re-grasps are done to collect enough data to identify the item and its pose with a Hidden Markov Model (HMM) and a cloth simulator. The method we present here targets this particular problem: making more informed grasps so as to shorten the initial data collection process.

Regarding the non-robotics computer vision literature, recognition and detection of deformable objects has attracted much less attention than those of their rigid counterparts. Furthermore, in the context of current object detection literature, the label *deformable object* often refers to those that have some degree of articulation, like people or animals, as in [5], rather than to highly flexible objects such as cloth.

Consequently, in this work we investigate if the well-known Bag of Features (BoF) [6] image representation is

suitable for this type of very deformable objects. It is known in computer vision literature that the geometry-less nature of BoF based methods makes them work better for flexible objects than template matching or pictorial structures methods [7]. We propose a method for detection of cloth parts that can be used prior to the first manipulation attempt by a laundry handling robot; in particular we have started with the detection of polo shirt collars (see Figure 1).

For detection, we base our method on a classical sliding window approach [8]. Different variants of this method are typically used in detection algorithms e.g. the ones participating in the yearly Pascal Visual Object Challenge detection competition. We have drawn inspiration from the methods proposed by Harzallah et al. [9] and Aldavert et al. [10].

Since our robot manipulation system uses a Kinect sensor for data acquisition, we evaluate the performance of the Geodesic-Depth Histogram (GDH) descriptor, in addition to that of the photometric only Scale Invariant Feature Transform (SIFT) [11].

II. Proposed Method

As said in the introduction, the objective of the method we propose is to detect an informed initial grasping point, which can be useful for an end-to-end cloth handling system like the one of Cusumano-Towner et al. [4], for example to shorten the series of re-grasps necessary to verify that the cloth is in a suitable state for the planning algorithm. We attempt to use state of the art computer vision techniques to detect the relevant grasping points from the very beginning, while the object is still laying on the table/surface. For this we propose a vision and depth based detection method, consisting of a three layer architecture. At this stage, and as done in related work, we are not considering the problem of background subtraction as a considerable body of work is already addressing it. We assume a segmentation method able to precisely select the cloth object is provided. In our case we accomplish this using a simple color threshold.

A. Appearance and depth local features

Our detection method is based on appearance and depth information, obtained from the Kinect image. In order to obtain the geometric information, we use the Geodesic-Depth Histogram (GDH), which captures the joint distribution of geodesic distances and depths within the patch. It is inspired by the Geodesic-Intensity Histogram, originally introduced by Ling and Jacobs [12] for describing deformable image patches.

Let us consider a patch $\mathcal{P}$ in the image, centered on a point of interest p, that in our case corresponds to every point of a grid that densely covers the image. Each point $p_i \in \mathcal{P}$ has an associated depth value d_i obtained from the Kinect camera. We then compute the histogram of p as follows:

- We initialize the histogram, by splitting the joint space of geodesic distances and depth into a discrete number of intervals. In our experiments we used a 11×8 discretization.
- For each $p_i \in \mathcal{P}$, we compute the geodesic distance g_i with respect to p, using the Fast Marching algorithm [13].
- We then fill the bins of the histogram with each pair (d_i, g_i) of depth and geodesic distance values.

The descriptor of p is finally built by concatenating the value of all the bins of the histogram, resulting in a 88-dimensional vector.

Regarding the texture information, we use the well-known Scale Invariant Feature Transform (SIFT). This descriptor divides a local patch around the interest point p in 16 sub-regions, and computes a 8-bin histogram of the orientations of the gradient for each sub-region, weighted by its corresponding magnitude and a Gaussian applied at the center of the patch. In order to reduce the aliasing in the orientation, trilinear interpolation is used to distribute gradient samples across adjacent bins of the histograms. Next the histograms are concatenated, yielding a 128 dimensional descriptor. To reduce the influence of non-affine illumination changes, the normalized descriptor is thresholded at 0.2 and re-normalized.

Both types of features are quantized using visual vocabularies learned with K-Means from a large training database of descriptors. A Bag of Features descriptor can be constructed by accumulating in a histogram all the visual words present in a local neighborhood defined by a bounding box.

B. Detection probability map

With BoF descriptors constructed from positive and negative training bounding boxes, a logistic regression model is trained using LIBLINEAR to obtain the posterior probability of the polo collar being present in a given bounding box. The probability of a bounding box containing a collar (class C_1) given a BoF descriptor x can be computed as:

$$p(C_1|x) = \frac{1}{1 + e^{w^T x}} \tag{1}$$

where w are the parameters of the model, learned minimizing the following expression:

$$\min_w \left(\frac{1}{2} w^T w + C \sum_{i=1}^{N} log(1 + e^{-y_i w^T t_i}) \right) \tag{2}$$

where C is the regularization parameter (adjusted by cross-validation), t_i stands for the i_{th} training example and y_i is its corresponding label.

Positive samples are the annotated bounding boxes in the training set, and negatives are random bounding boxes, sampled from the cloth area, that do not have more than 50% overlap with the annotated bounding box according to the Jaccard index:

$$I_{Jaccard} = \frac{area(B_n \cap B_{gt})}{area(B_n \cup B_{gt})} \tag{3}$$

where B_n is the generated negative bounding box and the B_{gt} is the ground truth one.

In the first layer of the architecture (corresponding to steps b and c of Figure 2) the logistic regression model is used in

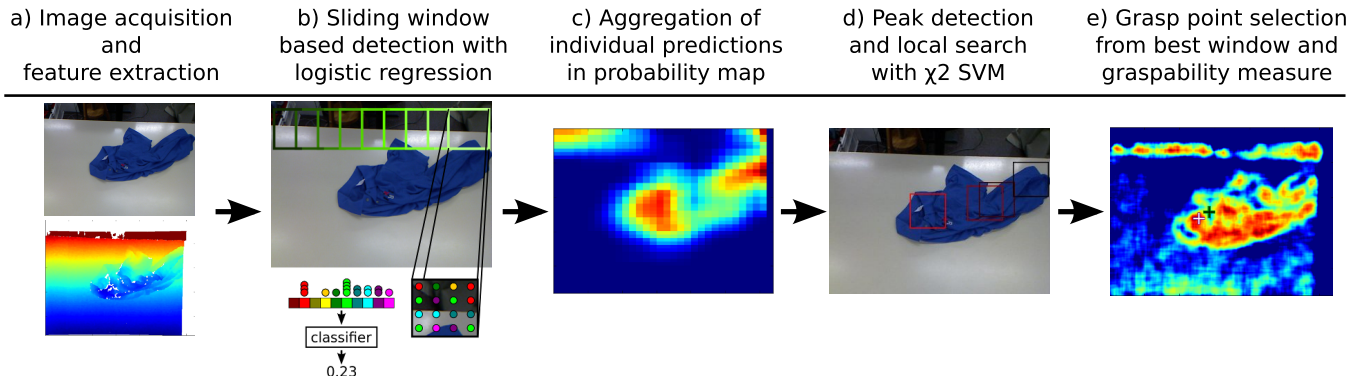


Fig. 2. Schema of the proposed method. Steps b and c correspond to the first layer of the approach as described in the text. Step d corresponds to the second layer, and step e to the third. In the image of step d, reddish color of the bounding box indicates more confidence in the detection. In the image of step e, the black cross indicates the initial grasping point, and the white cross the grasping point after searching in the "wrinkledness" image.

a sliding window approach covering the whole image, with different sizes and shapes of bounding boxes drawn from the distribution of those annotated in the training set. Next, similarly as it is done in [10], the probabilities of all windows are combined in a probability map of the presence of a collar (see for instance the middle image in Figure 3). Local peaks of this probability map are then selected and passed to the second stage of the architecture.

C. Detection refinement

A linear method like logistic regression has the advantage of being fast to apply at test time, but its performance is sometimes limited. A type of classifier with more capacity (potentially infinite), and specifically designed for histograms, is the Support Vector Machines with the χ^2 extended Gaussian kernel [14]:

$$\chi^2(x,t) = exp(-\gamma \sum_j \frac{(x_j - t_j)^2}{x_j + t_j}) \quad (4)$$

where γ is the inverse of the average of the χ^2 distance between the elements on the training set.

In the second layer (corresponding to step d of the Figure 2), for each selected candidate point, we cast a set of windows of different shapes and offsets with respect to the original point. Next, the score assigned by a χ^2 kernel SVM is used to rank these new windows, and only the highest ranked window for each candidate point is accepted. In practice we are conducting a local search around the most probable locations of the collar with a more expensive but reliable classifier.

D. "Graspability" measure

Finally, in the third layer (corresponding to step e of Figure 2), a graspable point is selected based on depth information. The most commonly used way to ensure that a point is graspable is by selecting the one that maximizes height. Although this measurement works well in controlled environments, it is no guarantee that the point will be easily graspable by a robot hand having limited precision. Instead, we propose a "wrinkledness" measure, that will give more importance to regions forming a pyramidal or conic structure, easily graspable by a robot hand. We define this measure as the entropy in the distribution of the orientation of the normal (expressed in spherical coordinates) in a local region.

The normal vector is computed for every point of the depth image using the information of its neighbors, and converted to spherical coordinates:

$$(\phi, \theta) = \left(arccos\left(\frac{z}{r}\right), arctan\left(\frac{y}{x}\right)\right) \quad (5)$$

where ϕ is the inclination and θ is the azimuth, (x, y, z) are the normal vector coordinates, and r is the radius, defined as:

$$r = \sqrt{x^2 + y^2 + z^2}. \quad (6)$$

Then, a bi-dimensional histogram of (ϕ, θ) pairs is constructed for a local region around each point (31 pixel side in our experiments) and its entropy is computed as:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i) \quad (7)$$

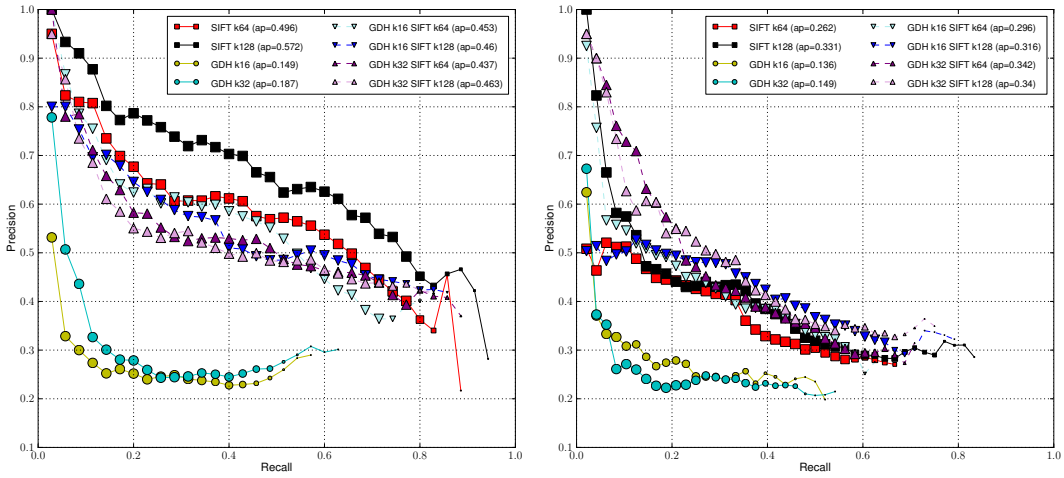
where X is the n-bin angle orientation histogram, and x_i is the ith bin. We used $n = 64$ in our experiments.

Entropy measures how much information exists in a message or distribution, or alternatively, how "predictable" it is. In our context, it directly tells us the amount of support of the distribution concentrated in high probability peaks or, equivalently, how much of the surrounding area of the point has normals aligned in the same orientation i.e. a flat surface or a combination of a few flat surfaces. See the rightmost image in Figure 3 for an example response image of this measure.

Our environment is calibrated, therefore we relate the 3D points of the camera with positions of the robot arm. Once the best grasping point is determined we can simply perform an open loop grasping to this point.

III. Evaluation Results

We have acquired a dataset of 194 images of various polo shirts with a Kinect camera, with large variations in pose. The images are of 640×480 pixels, the best resolution



(a) Blue polo subset

(b) Others subset

(c) Mixed subset

(d) No X^2 SVM

Fig. 4. Precision-recall curves of test with the three image subsets for different features and vocabulary sizes. The curves are the average of 10 complete runs of the method (training/testing). Size of the marker indicates proportion of the runs which reached the corresponding recall level. The legend also displays mean average precision. Figure (d) corresponds to a test with the blue polo subset comparing the results with and without the second layer of the architecture, i.e. not refining the peaks found from the probability map with the χ^2 kernel SVM classifier.

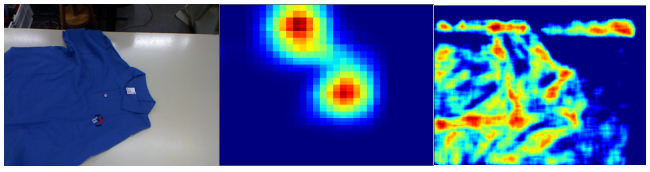


Fig. 3. From left to right: original image, detection probability map and "wrinkledness" measure.

offered by the camera when combining with depth. To train the classifiers, we have used only images of the blue polo, like the one that can be seen in the first column of Figure 3.

In addition to the blue polo, we also used other shirts to evaluate how well the models acquired in one piece of cloth could translate to similar ones. Furthermore, unrelated cloth pieces were added to test the robustness of the method to false detections. Various example pictures of the dataset can be seen in Figure 6[1]. Note the variability in appearance of the collar due to the flexibility of the textile material. For some images, it is even difficult for a human to determine where the collar is. We have divided our dataset in three subsets:

- **Blue polo**: This subset contains 117 images with the blue polo only. We split it in a training set (70%) and a test set (30%). The training set is used consistently for all experiments.
- **Others**: This subset is composed of images showing shirts other than the blue polo, and is used to evaluate how well the previously trained classifiers generalize to other similar, but different in appearance, cloth items. It contains 48 images, with one annotated collar per image.

[1]More result images can be found in the project web-page `http://www.iri.upc.edu/research/webprojects/intellact/planning/wrinkledGrasping.php`

- **Mixed**: The blue polo appears mixed with other cloth items. The collars of other shirts present in the images have also been annotated. This subset contains 29 images with a total of 48 annotated collars.

Since our objective is to find a good grasping point in the collar of the cloth, the evaluation measure we used in this work is the percentage of grasp point selections that fall inside its corresponding manually annotated bounding box. To account for randomness when sampling negative bounding boxes, all results reported are the average of ten complete runs of the method (training and testing).

Figure 4.a shows the average precision-recall curves for a 10 repetitions experiment using the *blue polo* subset. Precision corresponds to the percentage of correct detections, and recall to the percentage of annotated objects that have been detected. The precision-recall curve plots the precision obtained at each recall level. As can be seen, the SIFT features obtained the best results. In the other extreme, the GDH features attained lower precision and recall levels, and the combination of both features obtained results comparable to the SIFT ones. It can also be observed that the limited size and variability of the training set favors small visual word vocabularies, since larger ones introduce aliasing to the quantized feature space. Interestingly, experiments with the *others* subset (See Figure 4.b) show a decline in the performance of the SIFT features, while the GDH and the combined features attain similar results. Since SIFT features rely strongly on appearance, it seems reasonable that they perform worse when changing the object. On the other hand, the shape of the deformation may be expected to be similar. This suggests the value of the depth features for handling variation and generalizing to more types of clothes. When using the *mixed* subset all methods attain lower average precision levels, as can be seen in Figure 4.c. This degradation is not surprising, since we are dealing with a significantly more complex scenario. The good performance of the SIFT features in the highest recall levels is most likely due to the presence of the blue polo in several images of this subset.

The results shown in Figure 4.d illustrate the advantage of adding the second layer to our algorithm. The performance using windows generated only from the peaks of the probability map causes a drop of up to 20% in the average precision.

Figure 5 shows another interpretation of the results: it displays the percentage of true positives found looking only at each of the $N = 5$ top ranked windows of every image in a stacked bar plot for all the feature combinations and subsets. This interpretation is especially relevant for robot manipulation, since for a given image a decision must be taken regarding where (in 3D space) to attempt the grasp action, and this is exactly the window with the maximum score in our approach. As can be seen in the figure, the first choice (lower part of the stacked bar plot) was correct around 70% of the times for the *blue polo* subset, and about 60% for the *others*. The *mixed* subset attains a more modest 30%. Note that if there is no collar present in the image, a window would still be selected, even if it had a low classifier

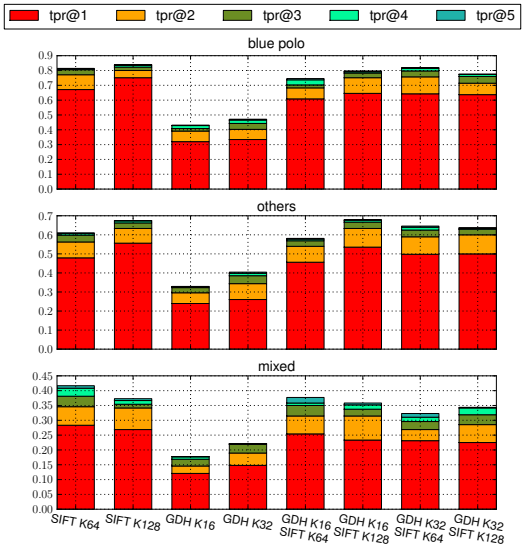


Fig. 5. Percentage of true positives considering only the $N = 5$ windows with the highest classifier score per image. Each stacked bar plot shows the fraction of true positives captured for each of the five best classified windows. Please notice the difference in scale in the vertical axis between the different subsets. The term *tpr* stands for true positive rate.

score. Therefore, this analysis must be taken with caution since, contrarily to precision-recall curves, it is not taken into account that at a certain classifier score, the grasp action would be inhibited. Therefore, some of the hits exhibited in this graph could have been filtered by this threshold on the classifier score.

We have used the learned classifiers in a real manipulation scenario. The experimental setup includes a Kinect camera rigidly attached to the environment, and a WAM arm with a three-fingered Barret hand (See Figure 1). For simplicity we used a predefined grasping position of the fingers, instead of adapting the position to the current state of the cloth object. Once the grasping point has been determined, the robot executes an approaching motion and performs the grasp of the object. We put in correspondence the two coordinate systems (i.e. camera and robot) using a common hand-eye calibration method.

IV. Conclusions

Usually, in robotic textile manipulation literature the perception component is reduced to the simplest possible system that fulfills the desired requirements in a controlled environment. We feel that, on the contrary, it is important to develop methods capable of properly perceiving and modeling cloth objects in order to efficiently execute the desired manipulation tasks. Therefore, in this work we proposed a method for the difficult task of detecting a part of a very flexible object using computer vision techniques on images enhanced with depth information, and we have shown its applicability in a real manipulation scenario. We have found that the SIFT appearance descriptors have a surprisingly good performance, especially in detecting deformed versions of the

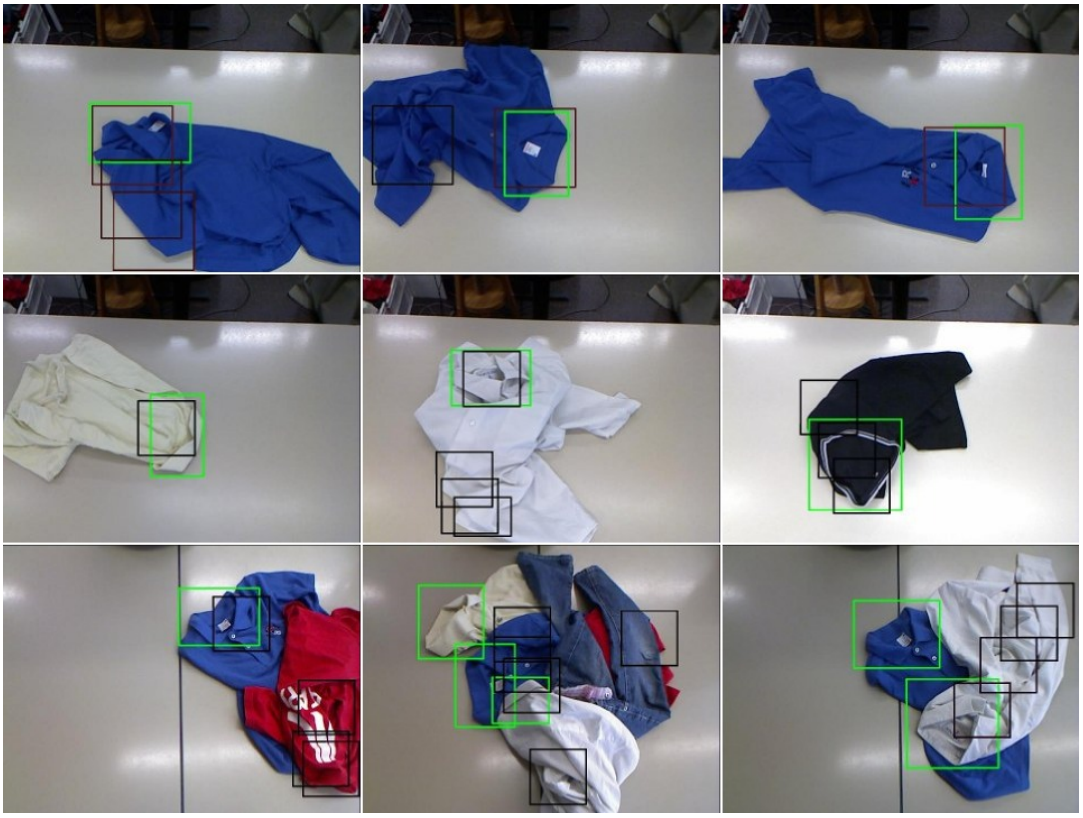


Fig. 6. Example detection images. The rows correspond to the *blue polo*, *others* and *mixed* subsets, in this order. Green boxes correspond to ground truth annotations. Notice the top-left and the middle-left examples, that show accurate detections of wrinkled collars.

same object (we achieved around 70% correct detections with the first classified hypothesis of each image for this task), but combining it with GDH shape descriptors allowed to improve the generalization capabilities of the method.

Future work includes testing the method with other cloth parts (such as the sleeve of a sweater, the hips of a pair of pants or the ankle of a sock), and evaluating alternative 3D descriptors, such as HKS [15] or DaLI [16].

REFERENCES

[1] B. Willimon, S. Birchfield, and I. Walker, "Classification of Clothing using Interactive Perception," in *Proc. IEEE International Conference on Robotics and Automation*, pp. 1862–1868, 2011.

[2] P. Wang, S. Miller, M. Fritz, T. Darrell, and P. Abbeel, "Perception for the Manipulation of Socks," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4877–4884, 2011.

[3] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel, "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *Proc. IEEE International Conference on Robotics and Automation*, pp. 2308–2315, 2010.

[4] M. Cusumano-towner, A. Singh, S. Miller, J. F. O'Brien, and P. Abbeel, "Bringing Clothing into Desired Configurations with Limited Perception," in *Proc. IEEE International Conference on Robotics and Automation*, (Shangai, China), pp. 3893–3900, 2011.

[5] M. Pedersoli, A. Vedaldi, and J. Gonzalez, "A Coarse-to-fine approach for fast deformable object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1353–1360, 2011.

[6] G. Csurka, C. R. Dance, L. Fan, C. Bray, and J. Willamowski, "Visual Categorization with Bags of Keypoints," in *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[7] O. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman, "The Truth About Cats and Dogs," in *Proc. International Conference on Computer Vision*, 2011.

[8] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518, IEEE Computer Society, 2001.

[9] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Proc. IEEE International Conference on Computer Vision*, pp. 237–244, 2009.

[10] D. Aldavert, A. Ramisa, R. Toledo, and R. L. D. Mantaras, "Fast and Robust Object Segmentation with the Integral Linear Classifier," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1046–1053, 2010.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] H. Ling and D. W. Jacobs, "Deformation invariant image matching," in *Proc. IEEE International Conference on Computer Vision*, pp. 1466–1473, 2005.

[13] J. Sethian, "A fast marching level set method for monotonically advancing fronts," *Proceedings of the National Academy of Sciences*, vol. 93, no. 4, pp. 1591–1595, 1996.

[14] J. Zhang, M. Marszaek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *International Journal of Computer Vision*, vol. 73, pp. 213–238, Sept. 2006.

[15] A. Bronstein, M. Bronstein, L. Guibas, and M. Ovsjanikov, "Shape google: Geometric words and expressions for invariant shape retrieval," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 1, p. 1, 2011.

[16] F. Moreno-Noguer, "Deformation and illumination invariant feature point descriptor," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1593–1600, 2011.